Implementación de Data Stream Mining

Esteban Schab[1], Ramiro Rivera[1], Luciano Bracco[1], Facundo Coto[1], Juan Manuel Ríos[1], Carlos Casanova[1], Patricia Cristaldo[1], Anabella De Battista[1], Norma Edith Herrera[2]

[1] Departamento Ingeniería en Sistemas de Información
Facultad Regional Concepción del Uruguay, Universidad Tecnológica Nacional
Entre Ríos, Argentina
{schabe, riverar, braccol, cotof, riosj, casanovac, cristaldop, debattistaa}@frcu.utn.edu.ar

[2] Departamento de Informática, Universidad Nacional de San Luis, San Luis, Argentina
nherrera@unsl.edu.ar

Resumen. Desde hace décadas las organizaciones utilizan información histórica propia para construir data warehouses y, mediante la aplicación de técnicas de descubrimiento de conocimiento, descubrir patrones que guíen la toma de decisiones. Actualmente, es una oportunidad para las organizaciones tomar decisiones en tiempo real basadas en información que puede provenir de múltiples fuentes, con diversos formatos y que se genera a gran velocidad. Como respuesta a esta necesidad surge Data Stream Mining (DSM), un subárea específica de la Minería de Datos definida como el proceso de extraer conocimiento en estructuras de datos continuas y con rápidas transiciones [1]. Dicho análisis aporta a las organizaciones visibilidad del negocio y de sus clientes en tiempo real y les permite responder ágilmente ante los cambios. En este trabajo se presenta la vinculación del GIBD de la UTN-FRCU con la empresa *Sidesys IT Solutions* [2] con el objetivo de implementar Data Stream Mining en la empresa.

1 Caracterización General del Proyecto

1.1 Instituciones y Empresas Participantes

- *Sidesys IT Solutions*: empresa de tecnologías de la información dedicada al diseño, desarrollo e implementación de soluciones innovadoras orientadas a la mejora de la experiencia de clientes.
- *GIBD*: Grupo de Investigación en Bases de Datos (GIBD) de la Facultad Regional Concepción del Uruguay, Universidad Tecnológica Nacional.

1.2 Personas Participantes

Nombre	Rol	Institución
Juan Manuel Ríos	Director de Operaciones	Sidesys
Facundo Coto	Desarrollador / Investigador	Sidesys / GIBD
Luciano Bracco	Investigador	GIBD
Ramiro Rivera	Investigador Alumno	GIBD
Esteban Schab	Docente Investigador	GIBD
Carlos Casanova	Docente Investigador	GIBD
Patricia Cristaldo	Docente Investigadora	GIBD
Anabella De Battista	Co-directora GIBD	GIBD
Norma Herrera	Directora GIBD	GIBD

Tabla 1: Listado de participantes del proyecto

1.3 Tipo de Interacción

Colaboración en I+D	X
Asistencia técnica de investigadores a empresas	X
Comercialización de resultados de I+D	
Desarrollo de currícula y clases en conjunto	
Emprendedorismo (start-up, spin-off)	
Otro. Especificar:	

Tabla 2: Tipo de interacción universidad-empresa

2 Detalles de Ejecución del Proyecto

2.1 Objetivos

Diseño y desarrollo de una solución para el procesamiento de streams de datos provenientes de un sistema de gestión de turnos desarrollado por *Sidesys IT Solutions*.

2.2 Actividades Realizadas

- *Actividad 1:* Análisis de requerimientos y definición de parámetros a observar (valores que resultan interesantes analizar dentro de los streams de datos, para mejorar el proceso de toma de decisiones). Estudio del modelo de negocio. Actividad realizada en conjunto con la empresa.

- *Actividad 2:* Investigación y selección de herramientas para el desarrollo de la arquitectura de procesamiento de streams de datos. Diseño de arquitectura en base a estas herramientas y a los requerimientos definidos.
- *Actividad 3:* Simulación del funcionamiento del negocio en conjunto con el de la herramienta diseñada mediante una carga de trabajo tomada de una base de datos histórica anonimizada. Dicha simulación permite estudiar el comportamiento del negocio y establecer niveles de servicio para la configuración de los prototipos de la solución.
- *Actividad 4:* Construcción de prototipos de la solución en base a: los requerimientos definidos, los parámetros obtenidos de la simulación, el análisis de los datos anonimizados.

2.3 Origen de los Fondos

- *Fondos de actividad 1, 2, 3 y 4:* las actividades involucradas en este proyecto se han financiado de la siguiente manera: personal del GIBD a través del financiamiento recibido como proyecto homologado de la Universidad Tecnológica Nacional, sueldos de sus investigadores y becas en el caso de alumnos investigadores; en el caso del personal de la empresa afectado a este proyecto, las actividades se han financiado con los sueldos correspondientes.

3 Resultados del Proyecto

3.1 Resultados de cada Actividad

- *Resultados de actividad 1:* Detalle de requerimientos y parámetros a observar, modelo de negocios.
- *Resultados de actividad 2:* Diseño de la arquitectura a partir de las herramientas seleccionadas.
- *Resultados de actividad 3:* Prueba de carga de trabajo para cada componente de la arquitectura. Niveles de servicio mínimo a cumplir, necesarios para la validación del producto.
- *Resultados de actividad 4:* Prototipos funcionales que permiten realizar pruebas e iteraciones de mejora.

3.2 Evaluación de los Resultados y Lecciones Aprendidas

Los datos generados por el producto principal de *Sidesys IT Solutions*, una solución integral destinada a la gestión de flujo de personas para mejorar la eficiencia del proceso de atención al público, resultan la materia prima fundamental para realizar procesamiento de streams de datos, mediante el monitoreo en tiempo real de los niveles de servicio e indicadores clave de negocio, a través de herramientas operativas y de gestión como dashboards, alarmas y notificaciones que faciliten la toma de decisiones en tiempo real. La empresa cuenta actualmente con un departamento de I+D, sin embargo no le resulta viable desarrollar un proyecto de esta magnitud, por lo que se visualizó

como una oportunidad la vinculación con un grupo de investigación, ya que le permite compartir el esfuerzo que implica el estudio y adaptación de nuevas tecnologías a sus productos. Para el grupo de investigación representa la posibilidad de realizar transferencia de nuevas tecnologías a la industria y de contar con datos reales para el estudio de las mismas.

En la gestión de las actividades del proyecto se empleó una metodología ágil basada en Scrum [3] y CRISP-DM [4], que se espera formalizar como una propuesta de metodología ágil para la gestión de proyectos de ciencia de datos. Inicialmente, se realizó el análisis de una base de datos históricos anonimizados y luego, con los resultados obtenidos, la simulación del funcionamiento del negocio en conjunto con el de la herramienta diseñada. Dicha simulación permite probar el funcionamiento y calcular niveles de servicio para la configuración de los prototipos de la solución. En paralelo se avanza con la implementación de prototipos de la solución en base a los requerimientos definidos y a los niveles de servicio mínimo a cumplir.

Como lecciones aprendidas se puede mencionar que al inicio del proyecto se tomaron ciertas decisiones que luego condicionaron el proceso de desarrollo de los prototipos. Una de estas decisiones fue la de utilizar alguna forma de virtualización para poder generar entornos de desarrollo y prueba replicables de manera automatizada, siguiendo las prácticas y filosofías DevOps para los proyectos Agile [5]. En esa etapa se planteó la disyuntiva “Máquinas Virtuales vs Contenedores” [6, 7], optando por la primer opción. Sin embargo este camino no resultó adecuado, ya que si bien los integrantes del proyecto conocían las herramientas de virtualización, se presentaron varios inconvenientes en la configuración entre las máquinas virtuales. Finalmente se propuso el uso de virtualización mediante contenedores Docker [7], solución que se recomienda para el despliegue de cualquier arquitectura que implique requerimientos de comunicación y sincronización entre varias herramientas. Entre las ventajas de Docker se destacan que pone a disposición un ecosistema de tecnologías de contenerización y una plataforma para desarrollar, compartir y desplegar aplicaciones de manera rápida y predecible, con una amplia comunidad de soporte.

Referencias

1. Khan, L., Fan, W.: Tutorial: Data Stream Mining and Its Applications. In: Lee, S., Peng, Z., Zhou, X., Moon, Y.-S., Unland, R., and Yoo, J. (eds.) Database Systems for Advanced Applications. Springer Berlin Heidelberg, Berlin, Heidelberg (2012).
2. Sidesys IT Solutions, http://www.sidesys.com/.
3. Scrum Guide | Scrum Guides, https://goo.gl/YfRLzP.
4. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: Step-by-step data mining guide.
5. Agerbak, A., Burchardi, K., Kok, S., Lebegue, F., Schmid, C.: Going All In with DevOps, https://goo.gl/LH6n98.
6. VirtualBox, https://www.virtualbox.org/.
7. Docker, https://www.docker.com/.